Apples and Oranges

A Methodological Framework for Basic Research into Audiovisual Perception

Hans-Joachim Maempel

Citation:
Maempel, Hans-Joachim (2019). Apples and oranges: a methodological framework for basic research into audiovisual perception. In S. Hohmaier (Ed.), *Jahrbuch 2016 des Staatlichen Instituts für Musikforschung Preußischer Kulturbesitz* (pp. 361–377). Mainz et al.: Schott.
Open Access version retrieved from https://dx.doi.org/10.14279/depositonce-6424.2

This publication is based on the project »Audio-visual perception of acoustical environments« within the framework of the research unit *Simulation and Evaluation of Acoustical Environments (SEACEN)*, funded by the German Research Foundation (DFG) and coordinated by the Audio Communication Group of the TU Berlin.

Correspondence to be sent to Dr. Hans-Joachim Maempel, Department for Acoustics and Music Technology | Studio Facilites and IT, Staatliches Institut für Musikforschung Preußischer Kulturbesitz, Tiergartenstraße 1, Berlin, Germany, maempel@sim.spk-berlin.de

Hans-Joachim Maempel

Apples and Oranges

A Methodological Framework for Basic Research into Audiovisual Perception

I Introduction

Our access to the everyday world and many arts is, inter alia, inherently audiovisual, as long as no hearing or sight impairment handicaps perception. This statement may be trivial, but the questions that it raises are fundamental, and answering them by means of the empirical sciences is challenging. To what extent are we influenced by the voice and the facial expression of a person? To what extent are the localization of objects and the temporal determination of events based on hearing and sight? To what extent do the reaction time and the decisions of drivers depend on optical and acoustic information? How are pictorial and musical aspects of a film conducive to suspense? Does the visual aspect of a musical performance contribute to its emotional impact? And what role do the acoustic and optical properties of the performance space play? In this paper, several methodological issues raised by the experimental investigation of such questions are discussed and methodological criteria are derived. Finally, a technical research tool adequate to these criteria is presented.

II Subject

Research on audiovisual perception and processing indicate that the mental representation of physical objects is normally based on, amongst others, both the auditory and the visual modality. This may be observed in various fields such as intensity rating,[1] localization,[2] motion perception,[3] event time perception,[4]

1 Joseph C. Stevens, Lawrence E. Marks: »Cross-Modality Matching of Brightness and Loudness«, in: *Proceedings of the National Academy of Sciences of the United States of America 2,* 1965, pp. 407–411; Barry E. Stein, M. Alex Meredith: *The Merging of the Senses*, Cambridge/MA 1993, pp. 15–19.

2 Garth J. Thomas: »Experimental Study of the Influence of Vision on Sound Localization«, in: *Journal of Experimental Psychology 28,* 1941, pp. 163–177; Ian P. Howard, William B. Templeton: *Human Spatial Orientation*, London 1966; Mark B. Gardner: »Prox-

synchrony perception,[5] perceptual phonetics,[6] quality rating,[7] construction of meaning,[8] and room perception.[9] There is also comprehensive work on the neurophysiological mechanisms of multisensory perception and integration.[10] Despite this, both research on auditory perception and research on visual perception have paid little regard to the interplay of the auditory and visual modalities for a long time. Thus, many of the aforementioned fields still lack comprehensive theories of audiovisual perception and corresponding research strategies. Rather, researchers have considered quite specific variables and applied experimental paradigms and methods that are circumscribed accordingly.[11] As a consequence, it is not always possible to connect the results of various experiments in a systematic manner. However, by pursuing general objectives such as considering strategic aspects, critically reviewing methodologies, devising consistent terminology, and developing an empirically founded model the significance of research projects on audiovisual perception could be enhanced quite considerably.

imity Image Effect in Sound Localization«, in: *Journal of the Acoustical Society of America 43*, 1968, p. 163.

3 Armin Kohlrausch, Steven van de Par: »Audiovisual Interaction in the Context of Multi-Media Applications«, in: *Communication Acoustics*, ed. by Jens Blauert, Berlin et al. 2005, pp. 109–138.

4 Ladan Shams, Yukiyasu Kamitani, Shinsuke Shimojo: »Visual Illusion Induced by Sound«, in: *Cognitive Brain Research 14*, 2002, pp. 147–152; Tobias S. Andersen, Kaisa Tiippana, Mikko Sams: »Factors Influencing Audiovisual Fission and Fusion Illusions«, in: *Cognitive Brain Research 21*, 2004, pp. 301–308.

5 Berta Luise Heide, Hans-Joachim Maempel: »Die Wahrnehmung audiovisueller Synchronität in elektronischen Medien«, in: *26th VDT International Convention, Leipzig, 2010*, ed. by Bildungswerk des Verbands Deutscher Tonmeister, Bergisch-Gladbach 2010, pp. 525–537, https://www.tonmeister.de/tmt/index.php?p=de2010congress_d3&pga=pe05#pe05 [11. 4. 2018].

6 John MacDonald, Harry McGurk: »Visual Influences on Speech Perception Process«, in: *Perception & Psychophysics 24*, 1978, pp. 253–257.

7 John G. Beerends, Frank E. de Caluwe: »The Influence of Video Quality on Perceived Audio Quality and Vice Versa«, in: *Journal of the Audio Engineering Society 47*, 1999, pp. 355–362.

8 Claudia Bullerjahn, Markus Güldenring: »An Empirical Investigation of Effects of Film Music Using Qualitative Content Analysis«, in: *Psychomusicology 13*, 1994, pp. 99–118; Annabel J. Cohen: »How Music Influences the Interpretation of Film and Video: Approaches From Experimental Psychology«, in: *Perspectives in Systematic Musicology*, ed. by Roger A. Kendall and Roger W. H. Savage, Los Angeles 2005, pp. 15–36.

9 Hans-Joachim Maempel, Matthias Jentsch: »Auditory and Visual Contribution to Egocentric Distance and Room Size Perception«, in: *Building Acoustics 20*, 2013, pp. 383–402.

10 For an overview see *The New Handbook of Multisensory Processing*, ed. by Barry E. Stein, Cambridge/MA et al. 2012.

11 For a similar statement see Kathleen E. Shaw, Heather Bortfeld: »Sources of Confusion in Infant Audiovisual Speech Perception Research«, in: *Frontiers in Psychology 6,* 2015, https://doi.org/10.3389/fpsyg.2015.01844 [17. 12. 2018].

III Towards a general working model

Against this background of patchy findings on audiovisual perception, the primary goal of research efforts in the applicable fields should be the formation of empirically founded models. A general working model based on two ontological realms would be apt to describe fundamental effects of complex (in terms of multidimensional) physical properties on complex perceptual features within and across the modalities. The general working model might be differentiated by taking into account specific (in terms of unidimensional) physical properties, physiological and neurophysiological processing stages, specific perceptual features, and be extended by introducing feedback loops (e.g. orientational reactions, top-down processes, etc.). The investigation of further factors (contexts, personal features, etc.) could yield supplementary information about the scope of the model.

In order to build such an empirically founded model efficiently, it appears to be useful to apply a funnel-shaped research strategy: an initial investigation into basic questions leads to the more specific ones (or prompts the inclusion of existing specific results). Examples of research questions (RQs) representing different levels of specificity are:

1 What are the proportional contributions of hearing and sight to (complex or specific) perceptual features?
2 Do the modalities interact, and if so, in what way?
3 What are the effect sizes along effect directions within and between modalities and ontological realms, and – more specifically – between certain physical, physiological, neurophysiological and mental processing stages?
4 What is the scope of the findings about the above questions regarding perceptual conditions (quality of stimulus presentation, context), personal features (socio-demographic features, expertise), stimulus type (speech, music, noise), and semantic content?

Taking research efficiency into consideration, the use of a corpus of interrelated data is desirable. To this end, data should be collected concurrently from different realms, modalities, as well as processing stages; also different design paradigms involving different bi-modal stimulation principles should be integrated. An integrative data collection provides the opportunity to reveal multilevel and complex relationships between perceptual features, and allows to retrace the perception process along a larger section of the transmission and processing chain.

IV Methodological considerations

a Ontological realms

As a prerequisite for modeling both two modalities (sound and vision) and two ontological realms (the physical and the perceptual), a clear factual and terminological distinction of their respective categories has to be made. Thus I refer to the relevant physical properties as *acoustic* and *optical* and to the respective perceptual features as *auditory* and *visual*. As a complement to the term *modality* that differentiates between system-specific processes in the perceptual realm such as hearing and sight, I apply the term *domain* to the physical realm for distinguishing processes that are based on sound and light (table 1).

		Realm			
		physical	perceptual		
Domain	acoustic	*properties*	*features*	auditory	**Modality**
	optical	*properties*	*features*	visual	

Table 1: Distinctions between ontological realms and between domains and modalities, respectively.

Consequently, according to the suggested taxonomy, the term *stimulus* denotes a mere physical condition or event. Within this taxonomy, collocations such as *stimulus modality* or *visual stimulus* do not make sense because they blur the boundary between the physical and the perceptual realm.

Similar inconsistencies are a problem also encountered in classical psychophysics. The expression of sensory quantities by means of physical quantities may meet the test criterion of reliability to a certain degree, but fails to meet construct validity by definition.[12] For example, we do not *hear* a frequency spectrum itself, but rather the corresponding perceptual features of timbre and eventually pitch. And in contrast to the psychophysical Mel scale,[13] the perception of

12 Cf. Nicola Döring, Jürgen Bortz: *Forschungsmethoden und Evaluation in den Sozial- und Humanwissenschaften*, Berlin/Heidelberg ^{5}2016, pp. 98–99.

13 Stanley Smith Stevens, John Volkman, Edwin Newman: »A Scale for the Measurement of the Psychological Magnitude Pitch«, in: *Journal of the Acoustical Society of America 8*, 1937, pp. 185–190.

pitch is at least two-dimensional along a linear dimension (pitch height) and a cyclic dimension (pitch chroma), and a clear line between pitch and timbre aspects can hardly be drawn.[14] The English language often does not make such differentiation of the ontological realms easy. For example, both the physical and the perceptual representation of a sonic event are called ›sound‹. By contrast, in German, the physical event itself is termed ›Schall‹, while its associated perceptual representation is termed ›Klang‹.

In everyday life we are normally exposed to stimuli from the two domains in question (and from further domains not discussed here). However, in certain situations (e.g. listening to music with closed eyes) and under certain experimental conditions (sec. IV g) our senses receive information from only one domain. So, a distinction is to be drawn not only between the domains as such, but also between multi-domain and single-domain conditions of perception.

b Processing stages

Both integrative data collection (sec. III) and the differentiation of effect directions (sec. IV d) require the introduction of transmission stages in the physical realm and processing stages in the perceptual realm – irrespective of the extent to which each stage is causally determined by the previous stage. Naturally, in the macroscopic physical realm deterministic relations are predominant whereas in the neurophysiological and perceptual realm probabilistic relations are prevalent, and super-additive or sub-additive effects occur.[15] A future extension of the outlined model should surely also take into account physiological and neurophysiological processing stages.

14 Cf. A. Bachem: »Tone Height and Tone Chroma as Two Different Pitch Qualities«, in: *Acta Psychologica 7*, 1950, pp. 80–88; Kazuo Ueda, Kengo Ohgushi: »Perceptual Components of Pitch: Spatial Representation Using a Multidimensional Scaling Technique«, in: *Journal of the Acoustical Society of America 82*, 1987, pp. 1193–1200; Janice Giangrande, Betty Tuller, J. A. Scott Kelso: »Perceptual Dynamics of Circular Pitch«, in: *Music Perception 20*, 2003, pp. 241–262; Paul M. Briley, Charlotte Breakey, Katrin Krumbholz: »Evidence for Pitch Chroma Mapping in Human Auditory Cortex«, in: *Cerebral Cortex 23*, 2013, pp. 2601–2610.

15 Cf. Nicholas P. Holmes, Charles Spence: »Multisensory Integration: Space, Time and Superadditivity«, in: *Current Biology 15*, 2005, R762–R764; Dora E. Angelaki, Yong Gu, Gregory C. DeAngelis: »Multisensory Integration: Psychophysics, Neurophysiology and Computation«, in: *Current Opinion in Neurobiology 19*, 2009, pp. 452–458; Barry C. Smith: »The Chemical Senses«, in: *The Oxford Handbook of Philosophy of Perception*, ed. by Mohan Matthen, Oxford 2015, p. 340.

c Modality-specificity

The first step in differentiating processing stages is to clearly distinguish between domain-/modality-specific and non-domain-/non-modality-specific properties/features (table 2): Acoustic and optical properties depend on physical properties that are not *domain*-specific: for example, material and structural properties. Correspondingly, auditory and visual features influence other perceptual features that are not *modality*-specific: for example, perceived material and structural features or aesthetic impressions.

Modality-specific or unimodal features such as *loudness* presume information from a specific sensory system and may be applied only to the system-specific sensation in a non-metaphoric, denotative manner. Within the processes of extracting increasingly abstract perceptual features, modality-specific features are low-level features as a rule. The terminology *low-level*, *mid-level* and *high-level feature* reflects the hierarchy of abstraction.[16]

In contrast, non-modality-specific features variably exploit information from several sensory systems resulting in high-level features. They may be further differentiated into intermodal and supramodal features: Typically, intermodal features are matching features such as perceived synchrony, and they result from comparing at least two modalities on the ground of their intersecting or coincident unimodal percepts; this process plays an important role in the development of perception in infancy.[17] Supramodal features such as perceived location, room size, aesthetic impressions or perceived emotions, however, appertain to superordinate processes to which different modalities may or may not contribute.[18] Naturally, both unimodal and supramodal features occur under both the single-domain and the multi-domain condition (sec. IV a), whereas the occurrence of intermodal features requires multi-domain input to the senses.

16 For a short explanation see Bria Long, Talia Konkle, Michael A. Cohen, George A. Alvarez: »Mid-Level Perceptual Features Distinguish Objects of Different Real-World Sizes«, in: *Journal of Experimental Psychology: General 145*, 2016, pp. 95–109.

17 Lorraine E. Bahrick, Robert Lickliter: »Perceptual Development: Intermodal Perception«, in: *Encyclopedia of Perception*, Vol. 2, ed. by E. Bruce Goldstein, Newbury Park/CA 2010, pp. 753–756.

18 Cf. Annie Pye, Patricia E. G. Bestelmeyer: »Evidence for a Supra-Modal Representation of Emotion from Cross-Modal Adaptation«, in: *Cognition 134*, 2015, pp. 245–251; Nicholas P. Holmes, Gemma A. Calvert, Charles Spence: »Multimodal Integration«, in: *Encyclopedia of Neuroscience*, Vol. 3, ed. by Marc D. Binder et al., Berlin et al. 2009, pp. 2457–2461.

Specificity	modality-specific	non-modality-specific	
Relation	unimodal	intermodal	supramodal
Necessary condition for the formation of the percept (the signified)	information from a specific modality is available	information from two or more specific modalities is available	information from any modality out of several appropriate modalities is available
Coverage of the term (the signifier)	specific modality	relation of specific modalities	all appropriate modalities
Examples of perceptual features	loudness, pitch, timbre, brightness, contrast, color	synchrony, synlocation	time, location, spatial dimension, material, aesthetics, emotions (perceived, felt), sense of presence

Table 2: Distinction between modality-specificities

Such non-modality-specific features are concomitants or results of the mental reconstruction of the physical world. This is why many material and structural properties/features appear in both the physical and the perceptual realm, and may be respectively described by means of the same terms and units. For example, the combination of the properties/features ›room, box-shaped, wood-paneled, height 3 m, width 4 m, length 7 m‹ is valid within the physical and the perceptual realm. In contrast to modality-specific features, perceived material and structural features may therefore be reliably compared with physical material and structural properties. Naturally, this does not hold for non-modality-specific features without a counterpart in the physical realm, e.g. evaluative features such as aesthetic impressions or individual states such as felt emotions.

d Effect directions

The categorization of properties and features by means of ontological realms, modalities/domains, and processing stages allows for the denotation of several effect directions, leading from independent toward dependent variables that may be empirically tested. I suggest the following denotations for the most important effect directions (figure 1):

Intra-modal effects (B → C, D → E), for instance, the effect of frequency spectrum (B) on the timbre/pitch perception (C) of a violin.

Cross-modal effects ([D]B → E, [B]D → C), for instance, the effect of color (D) on the perceived loudness (C) of a train emitting a constant sound pressure level [B];[19] the square brackets indicate a required additional stimulus which may be constant or – in a more complex test design – varied; in the special case of the investigation of genuine synaesthesia, however, the additional stimulus is unnecessary.

Trans-domain effects (A → B, A → D), for instance, the effect of surface material (A) on the reverberation time (B) and color spectrum (D) of a concert hall.

Trans-modal effects (C → F, E → F), for instance, the effect of perceived loudness dynamics (C) and colorfulness (E) on the aesthetic impression (F) of a TV program.

Supra-modal effects (A → F), for instance, the effect of physical distance on the perceived distance of a singer.

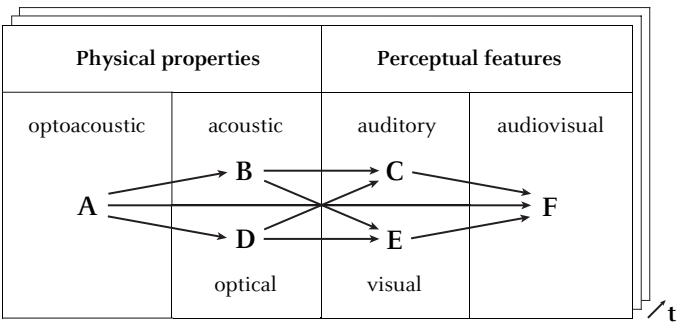

Figure 1: Basic effect directions

The modal specificity of perceptual features (sec. IV c) is closely related to the process of multimodal integration raising a crucial question known as the binding problem. Experimental paradigms in this field involve different times of domain-specific stimuli and responses, respectively.[20] With regard to those experiments the proposed outlined model might be orthogonally extended by the dimension of time. Thus, multiple sets of basic effect directions according to figure 1 would represent different time layers.

19 Hugo Fastl: »Audiovisual Interactions in Loudness Evaluation«, in: *International Conference on Acoustics, Kyoto 2004*, Vol. 2, Kyoto 2004, pp. 1161–1166.

20 For an overview see Sharon Zmigrod, Bernhard Hommel: »Feature Integration across Multimodal Perception and Action: A Review«, in: *Multisensory Research 26*, 2013, pp. 143–157.

e Operationalizing independent variables by means of dependent variables

Normally, experimental test conditions are varied by means of the selection or manipulation of (physical) stimuli or sensory organs. However, in some prior experiments, participants have been asked to rate a supra-modal feature, but to take into account information derived from only one modality (e.g. »participants were instructed to judge the emotion perceived auditorily«;[21] »*auditory expressivity*«[22]). Such instructions entrust test participants with the task of dissociating auditory and visual information. Which is to say, test participants are supposed to perceive a feature based on two modalities as if it relied on just one modality, despite the fact that such features rely upon audiovisual integrative processes that occur prior to (or in the course of) evaluation. Test participants might attempt to take on this challenge by means of directed attention; however, it is not clear whether participants are capable of suppressing both their conscious perception of excluded unimodal features, and the routing of unimodal information through preconscious audiovisual integrative processes. Alternatively, test participants might *ex post* try to assess the proportion of the non-modality-specific feature that is based on the demanded modality; however, the degree of validity and reliability involved in carrying out this cognitive task remains unclear. Moreover, it is not clear which of the two strategies might be pursued. Hence, compared to the experimental variation of (physical) stimuli, such instructions cannot guarantee a proper dissociation of auditory and visual information. They are useful at most for the investigation of directed attention.

f Interaction effects

General audiovisual cooperation is often referred to as ›audiovisual interaction‹.[23] In a colloquial sense, this term does not account for the different ways in which the senses cooperate (e.g. mutually supporting, mutually interfering, non-monotonic contributing), and at least from an empirical point of view its

21 Karin Petrini, Phil McAleer, Frank Pollick: »Audiovisual Integration of Emotional Signals From Music Improvisation Does not Depend on Temporal Correspondence«, in: *Brain Research 1323*, 2010, p. 144.

22 Jonna K. Vuoskoski, Marc R. Thompson, Eric F. Clarke, Charles Spence: »Crossmodal Interactions in the Perception of Expressivity in Musical Performance«, in: *Attention, Perception, & Psychophysics 76*, 2014, p. 598.

23 E.g. Riikka Möttönen, Mikko Sams: »Audiovisual Interaction«, in: *Handbook of Signal Processing in Acoustics*, ed. by David Havelock, Sonoko Kuwano and Michael Vorländer, New York et al. 2008, pp. 731–745.

use is frequently incorrect. In a methodological context, interaction effects occur whenever the effects of one independent variable depend upon a second independent variable (or variables) in a non-additive way.[24] In order to be able to reveal optoacoustic interaction effects, acoustic and optical variables have to be dissociated. Which is to say, acoustic and optical properties must be isolated as factors rather than as levels of one factor, and they must be varied mutually independently. Only then the two main effects and the interaction effect on each auditory, visual, and audiovisual variable can be independently quantified. The objective of quantification leads to the issue of design paradigms.

g Design paradigms

As indicated, some studies on the so-called audiovisual interaction treat hearing and sight as levels of factors rather than as factors themselves. Typically, such studies apply a factor called *presentation mode* or *modality* based on three levels (optical, acoustic and optoacoustic). Because the acoustic and optical stimuli are varied regarding their particular presence, this principle of variation may be called *co-presence (CP) paradigm*. It raises two issues.

Strictly speaking, the CP paradigm involves two sources of variation: the change between the presence and the absence of a certain stimulus domain, and the change between a single-domain and a multi-domain stimulus (sec. IV a). The need to dissociate these sources of variation is not only theoretically relevant: the value of a supra-modal dependent variable under the optoacoustic condition does not always equal the average of the values caused by the acoustic and optical conditions. So, the change between the single- and the multi-domain condition apparently also changes the basic mode of perceptual processing. Whereas a single-domain stimulus does not require a multimodal trade-off, a multi-domain stimulus plausibly does. Thus, acoustic and optical information may be weighted differently depending on the basic mode of perceptual processing. From a methodological point of view, the CP paradigm implies the confoundation of the two sources of variation at the cost of internal validity and differentiated theory formation. The sources of variation may, however, be dissociated at the analytical stage by testing the two single-domain levels (acoustic, optical) against each other and by separately testing the combined single-domain level (average of acoustic and optical) against the multi-domain level (optoacoustic). Since the respective set of a priori contrasts is orthogonal, multiple testing correction is not required.

24 Döring, Bortz: *Forschungsmethoden und Evaluation* (see note 12), p. 533.

Moreover, the CP paradigm is, at least formally, not apt to achieve the aim of testing for audiovisual interaction in a methodological sense (sec. IV f). Because no percepts normally occur in the absence of both acoustic and optical stimuli, in a full-factorial design one cell is empty; which is to say, the acoustic and the optical stimuli have not been mutually independently varied. Hence, the respective influence of the optical and acoustic domains may not be determined, let alone an optoacoustic interaction effect. In order for the acoustic and optical stimulus components to constitute mutually independent variables, as demanded for the determination of interaction effects, in a full-factorial design not their presence but their properties must be varied. This approach is often called *conflicting stimulus (CS) paradigm*. For example, a conflicting stimulus apt to investigate audiovisual localization might be realized by the synchronous presentation of the acoustic and optical components of a speaker in different locations. In contrast to the CP paradigm, the CS paradigm provides input to at

		Optoacoustic properties		
Perceptual features		1	2	3
Co-presence condition	A	av a	av a	av a
	OA	v av a	v av a	v av a
	O	v av	v av	v av

		Acoustic properties		
Perceptual features		1	2	3
Optical properties	1	v av a	v av a	v av a
	2	v av a	v av a	v av a
	3	v av a	v av a	v av a

			Acoustic properties		
Perceptual features		0	1	2	3
–	A		av a	av a	av a
Optical properties	1	v av	v av a	v av a	v av a
	2	v av	v av a	v av a	v av a
	3	v av	v av a	v av a	v av a

Table 3: Comparison of design paradigms: co-presence paradigm (left), conflicting stimulus paradigm (center), and integrated design (right).
Abbreviations: O = optical, OA = optoacoustic, A = acoustic, v = visual, av = audiovisual, a = auditory. Light grey shading indicates uni-domain stimuli, dark grey shading congruent two-domain stimuli, and no shading incongruent (conflicting) two-domain stimuli.
Due to the large number of cells, using within-subject factors might be practical. Analysis of variance (multivariate or univariate, with or without repeated measures) lends itself to data analysis in view of RQs 1 to 3 and, if applicable, RQ 4. RQs 3 and 4 might, however, require other, more specific designs and other statistical approaches, such as analysis of covariance, multiple regression analysis, or structural equation modeling.

least two modalities on all combinations of factor levels. Because the CP and CS paradigms speak to different conditions of optoacoustic experience and information-processing (see above), a comparison of both paradigms is of interest for theory formation. In view of integrative data collection (sec. III), merging both paradigms would be productive, as illustrated in table 3.

h Optoacoustic congruence

The quantification of the proportionate contribution of hearing and sight to perceptual features (sec. III, RQ 1) in particular requires compliance with further methodological criteria.

In reality, the acoustic properties of objects such as items, persons, rooms, environments, and whole scenes are indissolubly interrelated to their optical properties due to physical laws, and integrated perception processes are formed over the course of the corresponding long-termed experience. Under ecologically valid conditions, stimulus objects generally are optoacoustically congruent and thus ›sound as they look‹. Hence, the possibility of optoacoustic congruence is a prerequisite for ecologically valid applications of the CS paradigm. Accordingly, while the CS paradigm is designed to break down optoacoustic congruence in controlled measure, it still establishes the zero-conflict combination (perfect congruence) as a reference point from which other combinations of factor levels can be made to deviate.

The physical and perceptual determination of the degree of optoacoustic congruence is, however, only possible on the basis of several (i. e. complex) acoustic and optical properties. For example, illuminance and sound pressure level alone are not sufficient for this purpose. Because the determinability of congruence increases with the diversity of available acoustic and optical stimulus properties, perceptually relevant physical cues must not be removed from the acoustic nor the optical domain when presenting the experimental stimuli. I refer to the maintenance of perceptually relevant physical cues of stimulus objects as *rich cue condition*,[25] or in the ideal case as *full cue condition*.

Hence, a reliably determinable high optoacoustic congruence, as demanded by the CS paradigm, requires (1) a stimulus that is based on a real, naturally occurring object, and (2) its presentation under rich or full cue condition.

25 An established term, cf. e.g. Alessio Murgia, Paul M. Sharkey: »Estimation of Distances in Virtual Environments Using Size Constancy«, in: *International Journal of Virtual Reality 8*, 2009, p. 67.

i Simulation

Optoacoustically conflicting stimuli cannot be practically realized using real stimulus objects (sound and light sources and transmission systems, respectively) due to their natural optoacoustic congruence, as the above example of the speaker indicates (sec. IV g). Conflicting stimuli are incongruent by definition and must therefore be simulated. In order to maximize the ecological validity of the simulated stimuli in general and the optoacoustic congruence required by the zero-conflict combination in particular, the simulation has to be data-based (as opposed to numerically modeled) – that is, it has to display real instead of virtually designed objects. According to the rich or full cue condition, the simulation is furthermore required to be as transparent (i. e. physically correct) and immersive as possible. This may be achieved by technical features such as a transmission path with sufficiently high temporal and spatial resolution, the application of 3D audio and 3D video transmission methods, a largely nonrestrictive viewing/listening angle, a correct acoustic and optical projection geometry, the reproduction of real energetic conditions (sound pressure level, illuminance, dynamics), and shielding from distracting information. Of course, the question of whether the applied empirical methods of data collection yield results comparable for real stimuli and stimuli simulated in this way must be empirically looked into.

j Commensurability

Ensuring optoacoustic congruence is still not sufficient for an experimental appraisal of RQ 1 (sec. III). An internally valid quantitative comparison of the contributions made by the senses demands an identical range of stimuli. In other words, the respective variation of the acoustic and optical experimental stimuli must be quantitatively commensurable. A reasonable quantitative comparison of ranges requires their qualitative commensurability in turn. Naturally, this does not apply to the specific acoustic and optical properties expressed by means of different physical measures, for example, illuminance and sound pressure level. To apply the same numerical ranges to these two measures in an experiment would be to compare apples and oranges. Using an identical physical quantity, for example the power P, and applying an identical numerical range in the two domains would not offer a solution to this principle problem because the resulting acoustic or optical range would not be in line with power ranges of a real, naturally occurring stimulus object. Thus, the respective stimulus components were neither ecologically valid nor optoacoustically congruent

(sec. IV h). Qualitative commensurability of domain-specific properties may not be achieved in this way.

The commensurability problem may, however, be solved by relying on the qualitative commensurability of non-domain-specific properties, upon which the various domain-specific properties themselves depend. Given an experimental situation in which factor levels comprise the acoustic and optical components of several optoacoustically congruent stimulus objects, the range of the complex acoustic properties will largely correspond to the range of the complex optical properties. This is because on each level, the optical and acoustic properties derive from the same non-domain-specific properties such as physical materials, structures, and dimensions. Even though the complex variation of independent variables does not allow for an internally valid identification of unidimensional factors, it does well allow for the methodologically founded experimental appraisal of the basic RQ 1 (sec. III). Likewise, ensuring quantitative and qualitative optoacoustic commensurability requires a transparent simulation in order to meet the rich/full cue criterion (sec. IV h).

V Application: The Virtual Concert Hall

a Thematic background

The author of this article is currently investigating the above research questions 1 through 4 within the scope of an experimental research project on audiovisual room perception in which the various rooms themselves serve as stimulus objects. Presupposing a constant illumination, a room reveals its substantial optical properties in the form of light distribution and without any further contribution; which is to say, a room itself may be described as an optical stimulus from both a functional and a perceptual point of view. A room is not, however, a self-contained acoustic stimulus. Within the acoustic domain it is just a transmission system. Thus, the project requires sound sources capable of exciting the room's acoustics. So in addition to the above considerations, the investigation of room acoustics requires a differentiation between transmission system (room) and transmitted signal (rendition). In order to ensure the internal validity of the experiments, the rendition has to be held constant.

A research tool whose development was exclusively geared to the above-mentioned methodological criteria is the Virtual Concert Hall: an optoacoustic virtual environment for the presentation of artistic renditions in performance rooms such as concert halls, churches, and theatres. It allows for the mutually independent variation of the optical and acoustic components of both the ar-

tistic renditions and the spaces in which they are staged. Thus, although the factual stimulus objects are both renditions and rooms, it is possible to center the rooms as the stimulus objects of primary interest by holding the rendition constant across rooms. On the other hand, it also makes sense to vary the rendition with regards to the performance (interpretations), the performed content (works), and the type of content (music, literature) independently from the rooms, in order to improve the external validity of the experimental results regarding room perception.

b Technical implementation

Technical stages towards the realization of the Virtual Concert Hall include the acquisition of room properties, the production of artistic renditions, the merging of rooms and renditions, and the setup of a reproduction system – within both the acoustic and the optical domain, respectively.[26] The acquisition of room acoustic properties was carried out by recording binaural room impulse responses (BRIRs) for different azimuthal head orientations of a head-and-torso simulator.[27] The optical properties of the rooms were acquired in the form of stereoscopic full-panoramic images. The acoustic renditions were taken in an anechoic chamber applying poly-microphony and multitrack recording. The optical renditions were stereoscopically recorded in a green-box studio applying full playback. At the moment of reproduction, the acoustic renditions were embedded into the rooms by means of dynamic binaural synthesis, originally referred to as binaural room scanning.[28] This compensates for the head movements of listeners, resulting in a constant space-related localization across different head orientations. The resulting audio signal is reproduced by the use of an extra-aural headset and a DSP-driven power amplifier providing a linearized transfer function of the audio reproduction system.[29] The optical rendi-

26 Hans-Joachim Maempel, Michael Horn: »The Virtual Concert Hall: A Research Tool for the Experimental Investigation of Audiovisual Room Perception«, in: *International Journal on Stereo & Immersive Media 1*, 2017, pp. 78–98.

27 Alexander Lindau, Stefan Weinzierl: »FABIAN: An Instrument for Software-Based Measurement of Binaural Room Impulse Responses in Multiple Degrees of Freedom«, in: *24th VDT International Convention, Leipzig, 2006*, ed. by Bildungswerk des Verbands Deutscher Tonmeister, Bergisch-Gladbach 2006, pp. 621–625.

28 Ulrich Horbach, Attila Karamustafaoglu, Renato Pellegrini, Philip Mackensen, Günther Theile: »Design and Applications of a Data-Based Auralization System for Surround Sound«, in: *106th AES Convention, Munich 1999*, Preprint 4976, München 1999.

29 Vera Erbes, Frank Schultz, Alexander Lindau, Stefan Weinzierl: »An Extraaural Headphone System for Optimized Binaural Reproduction«, in: *Fortschritte der Akustik: Tagungsband der 38. DAGA 2012, Darmstadt*, ed. by Holger Hanselka, [Berlin], pp. 313–314.

tions were embedded frame-by-frame into the rooms by means of chroma-key compositing, the addition of shadows, and the correction of colors. The stereoscopic semi-panoramic high-resolution videos generated thereby were projected on a semi-cylindrical screen (d = 5 m; h = 2.8 m). In this manner, a large field of view (> 160°) and a reasonable physical resolution (4754 × 1872 pixels) corresponding to about 2.7 times the human eye's angular resolution could be realized. A secondary reproduction system based on an 85″ flat screen was also designed, providing much better angular resolution close perceptual threshold at a more restricted field of view. For experimental purposes, a test sequence control and an electronic questionnaire were also programmed.

c Features of the Virtual Concert Hall

Due to the combination of acquired room properties, produced content, and a state-of-the-art reproduction system, the Virtual Concert Hall allows for the presentation of identical artistic renditions under rich cue conditions in performance rooms with independently variable acoustic and optical properties. Thanks to several technical enhancements regarding the spatial resolution of the BRIRs,[30] the system latency,[31] the compensation of the headphone transfer function,[32] and the adaption of the interaural time differences to the individual listener,[33] the applied binaural synthesis system provides a highly plausible three-dimensional reproduction.[34] Sound sources and room reflections may be perceived omnidirectionally. Because the virtual scenes do not turn with head movements, they may be actively explored by the recipient. By acquiring further transmission systems and recording further content, the Virtual Concert Hall may be adapted for the data-based simulation of numerous scenes, allow-

30 Alexander Lindau, Hans-Joachim Maempel, Stefan Weinzierl: »Minimum BRIR Grid Resolution for Dynamic Binaural Synthesis«, in: *Acoustics '08, Paris,* [o. O.] 2008, pp. 3851–3856; Frank Schultz, Alexander Lindau, Stefan Weinzierl: »Just Noticeable BRIR Grid Resolution for Lateral Head Movements«, in: *NAG/DAGA 2009 International Conference on Acoustics, Rotterdam*, ed. by Marinus M. Boone, Berlin 2009, pp. 200–201.

31 Alexander Lindau: »The Perception of System Latency in Dynamic Binaural Synthesis«, in: *NAG/DAGA 2009* (see note 30), pp. 1063–1066.

32 Alexander Lindau, Fabian Brinkmann: »Perceptual Evaluation of Headphone Compensation in Binaural Synthesis Based on Non-Individual Recordings«, in: *Journal of the Audio Engineering Society 60*, 2012, pp. 54–62.

33 Alexander Lindau, Jorgos Estrella, Stefan Weinzierl: »Individualization of Dynamic Binaural Synthesis by Real Time Manipulation of the ITD« in: *128th AES Convention, London, 2010*, Preprint 8088, London 2010.

34 Alexander Lindau, Stefan Weinzierl: »Assessing the Plausibility of Virtual Acoustic Environments«, in: *Acta Acustica united with Acustica 98*, 2012, pp. 804–810.

ing for the experimental investigation of the diverse research questions outlined in sec. III.

Acknowledgements

This publication is based on the project »Audio-visual perception of acoustical environments« (MA 4343/1-1) within the framework of the research unit *Simulation and Evaluation of Acoustical Environments (SEACEN)*, funded by the German Research Foundation (DFG) and coordinated by the Audio Communication Group of the TU Berlin.